

Clicktionary: A Web-based Game for Exploring the Atoms of Object Recognition

D. Linsley* S. Eberhardt T. Sharma P. Gupta T. Serre
Cognitive Linguistic & Psychological Sciences Department
Brown University
Providence, RI, USA

*drew.linsley@brown.edu

Abstract

Understanding what visual features and representations contribute to human object recognition may provide scaffolding for more effective artificial vision systems. While recent advances in Deep Convolutional Networks (DCNs) have led to systems approaching human accuracy, it is unclear if they leverage the same visual features as humans for object recognition.

We introduce *Clicktionary*, a competitive web-based game for discovering features that humans use for object recognition: One participant from a pair sequentially reveals parts of an object in an image until the other correctly identifies its category. Scoring image regions according to their proximity to correct recognition yields maps of visual feature importance for individual images. We find that these “realization” maps exhibit only weak correlation with relevance maps derived from DCNs or image salience algorithms. Cueing DCNs to attend to features emphasized by these maps improves their object recognition accuracy. Our results thus suggest that realization maps identify visual features that humans deem important for object recognition but are not adequately captured by DCNs. To rectify this shortcoming, we propose a novel web-based application for acquiring realization maps at scale, with the aim of improving the state-of-the-art in object recognition.

1. Introduction

Advances in Deep Convolutional Networks (DCNs) have led to systems rivaling human accuracy in basic object recognition tasks [9]. While a growing body of work indicates this surge in performance carries concomitant improvement in fitting both neural data in higher areas of the primate visual cortex (reviewed in [35]) and human psychophysical data during object recognition [14], key differences remain. It has been suggested that processing depth

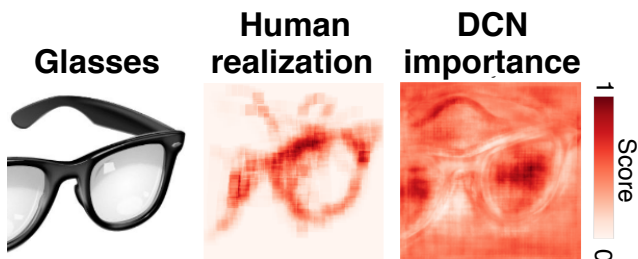


Figure 1. Humans and DCNs utilize different visual features for object recognition: “Realization” maps are derived from human observers playing *Clicktionary*, in which regions are scored according to their relevance for correct recognition and compared with feature importance maps obtained for DCNs.

achieved by our visual system arises through cortical feedback rather than through static processing stages [6]. It has also been shown that DCNs do not generalize well to atypical scenes, such as when objects are presented outside of their usual context [6, 24] or when features strongly diagnostic for recognition are absent [31]. This raises the possibility that DCNs may leverage entirely different visual strategies than humans during object recognition.

The most direct evidence for qualitatively different visual strategies used by humans and DCNs was demonstrated recently [31]. The authors located patches in several object images where human recognition was correct, but any further reduction of these patches (i.e., by cropping or down-sampling) led to incorrect responses. These were called the minimally reducible configurations of the images for recognition, or MIRC. Computer vision algorithms including DCNs failed to exhibit the same dramatic all-or-nothing recognition of MIRC as human participants. In other words, the object features visible in MIRC were more diagnostic of object category for humans than for DCNs. Although promising, this approach required about 14,000 participants to identify MIRC for 10 images, which in to-

tal yielded less than 200 MIRC's at varying resolutions and sizes. This makes it difficult to identify critical visual features in object images and seriously limits the applicability of this method to gather enough data for a more systematic analysis over many image exemplars and object categories.

To address these limitations, we created *Clicktionary*, a collaborative web-based game for identifying visual features that are necessary for object recognition. Pairs of human participants work together to identify objects: One player reveals diagnostic image regions while the other tries to recognize the object as quickly as possible. Amassing game-play data across many participants yields importance maps for individual images, in which each pixel is scored according to its contribution towards “realizing” an object’s category. This is illustrated in the middle panel of Fig. 1: the hotter the pixel, the more likely it caused recognition.

Here we compare the visual features used by humans, DCNs, and other computational models of vision. Using realization maps from *Clicktionary* we show that: (1) Realization maps derived from humans during object recognition are dissimilar to importance maps derived from DCNs and salience maps derived from attention models. (2) Using realization maps to guide DCNs towards object features that are diagnostic for humans leads to a small but significant improvement in recognition accuracy. (3) Motivated by these findings, we introduce clickme.ai, a web-based application for gathering realization maps at scale and measuring their impact on DCN performance in real-time.

2. Related Work

Behavioral studies: A central goal in vision science is to understand what features the human visual system uses to process complex visual scenes. Classic approaches to uncover the internal representations contributing to behavior include reverse correlation methods. These generally involve analyzing the statistical relationship between recognition and visual perturbations applied at different locations in a stimulus across many trials. While these methods have helped identify visual features that are diagnostic for faces and other synthetic object stimuli [26, 20], they typically require thousands of trials per subject to deduce these representations despite little shape variability for the studied object classes. Thus, they seem impractical for characterizing visual representations used for categorizing general classes of objects with higher variability in stimulus appearance, location, or lighting.

Another way of exploring visual feature importance is by recording eye fixations. Patterns of eye fixations represent observers’ efforts to center their high-acuity fovea on salient or task-relevant information, and capture diagnostic features in images [36, 8, 12, 13]. However, it remains difficult and costly to acquire large-scale eye tracking data, leading researchers instead to track computer mouse move-

ments during task-free viewing of images to estimate local saliency cues for fixations [18, 11].

Cognitive psychologists have classically used similarity judgments between image pairs as the gateway to studying visual representations. Recent work has compared similarity judgments derived from human judgments and representative DCNs and found good agreement between the two [22]. Related work has also evaluated the ability of DCNs to predict memorability [5] and typicality of individual images [17].

Computational models: A growing body of research focuses on understanding the nature of the visual features used by DCNs for object recognition. These methods fall into one of two groups. Sensitivity analyses use either gradient-based approaches [37, 28] or systematic perturbations of the stimulus to estimate local pixel-wise contributions of visual features to a classification decision [38]. Decision analyses such as layer-wise relevance propagation (LRP) estimate global pixel responsibility for the classification decision. Representative methods from both approaches are used here to derive importance maps from DCNs for comparison with those derived from humans.

Web-based games for data collection: There is a long history of leveraging the wisdom of the crowd through web-based applications to gather high-quality data for computer-vision studies. Closely related to our proposed *Clicktionary* game are the ESP game for identifying objects in real-world images [33] and the Peek-a-boom game for locating them [34]. In both of these games, participants work together to recognize an image of an object. We take particular inspiration from the mechanics of Peek-a-boom, where one participant in a pair reveals parts of an image to get the other participant to respond in some way about it. However, the mechanics of Peek-a-boom and *Clicktionary* are different in several key ways: (1) Because the goal of Peek-a-boom was to localize objects in images, the interface provided a large cursor for unveiling regions of low-resolution images. A higher-resolution interface was needed to study the importance of local visual features for object recognition. (2) Peek-a-boom players solve problems by sending each other visual “hints” about what should be located, and rate the accuracy of each others guesses. This is obviously likely to confound any interpretation of the results in terms of the diagnosticity of visual features. (3) More generally, there were no constraints in how Peek-a-boom players revealed images. This meant that players could adopt strategies that would interfere with the goals of the current study such as “salt and peppering” the screen with clicks or waiting long intervals between clicks.

Other web-based games have used similar interfaces as *Clicktionary* to explore various questions about human per-

ception. This includes a game for annotating image properties, where players take turns outlining important objects in real-world scenes and guessing their identities [30]. Similarly, visual Question Answering (VQA) was explored in a game where participants de-blurred parts of a scene image that were important for answering a question about it [4]. The game style has also been used to answer questions in biological and physical sciences, such as predicting protein structure [2] or neuronal connectivity [15].

Human-in-the-loop computing: Despite rapid advances in machine performance on vision tasks humans remain the gold standard, particularly when making decisions on images depicting atypical views or containing many occlusions. Researchers have found that human perceptual judgments in cases like these can augment the performance of models for vision. This has led to significant gains in face recognition and localization [25], action recognition [32], object detection and segmentation [21, 27].

3. Identifying features for object recognition

Clicktionary: Clicktionary was constructed to support the identification of visual features that cause recognition decisions in humans. Upon starting the game, players provided informed consent, read instructions, and were placed into a virtual waiting room. The waiting room contained a scoreboard listing the performance of the most successful teams to play the game. Our hope was to provide an incentive for players to compete with each other in order to collect the highest quality behavioral data possible. Participants were automatically paired in the waiting room¹. Once the game began, players collaborated over a series of rounds to name the category of an object image (Fig. 2).

Players accomplished this by alternating between two different roles over many game rounds. In each round, one player was the student and the other the teacher. The job of the teacher was to reveal areas of an image that were most informative for visual categorization. Teachers did this by clicking on the image and dragging their mouse, like a paint brush, to reveal its informative parts. Images were 300x300 pixels, and each revealed image patch was 18x18 pixels. These are hereafter referred to as “bubbles”. Teachers were instructed to choose their first click carefully because bubbles were automatically placed after the first one with a random interval between 50ms and 300ms. The center of each bubble was constrained to be within the radius of the previous one. Translucent blue boxes on the teacher’s image marked the image regions visible to the student.

In parallel, the student began each round viewing a blacked-out image and a text box requesting a guess of the

¹If no one else entered the waiting room within 120 seconds, players played against a DCN opponent. These data were not included in the current study.

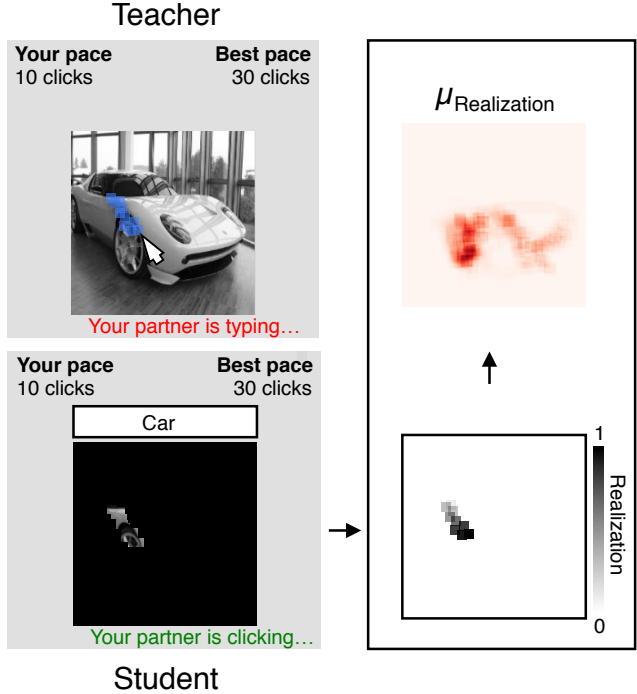


Figure 2. Overview of *Clicktionary*. Pairs of participants, one teacher and one student, played together to categorize objects. The teacher used the mouse to reveal image regions while the student performed a visual categorization task. Scoring revealed regions according to their temporal proximity to correct recognition yielded per-participant “realization maps”. Averaging across participants yielded realization maps that identified visual features causing recognition.

category of the object in the image. As the teacher bubbled in image regions, the corresponding locations of the blacked-out image were unveiled. Students were instructed to name the basic-level category of the object. For instance, the desired response for an image of a border collie was “Dog”. However, we also accepted subordinate-level category labels in order to expedite game play.

To provide players with incentive to work as quickly and efficiently as possible, their team’s performance versus the average performance of the top-10 teams was visible throughout the game. Team performance was measured as the number of image bubbles placed by teachers before students recognized the image. Although there was no explicit penalty for wrong answers, participants achieved better scores by avoiding them. If they finished the game in the top-10 they were congratulated and shown their ranking.

Incorrect guesses caused a red outline to appear around the image viewed by both student and teacher. Following a correct guess, the images were briefly outlined in green before participants began the next round. If the student could not figure out the object’s class, he pressed a skip button

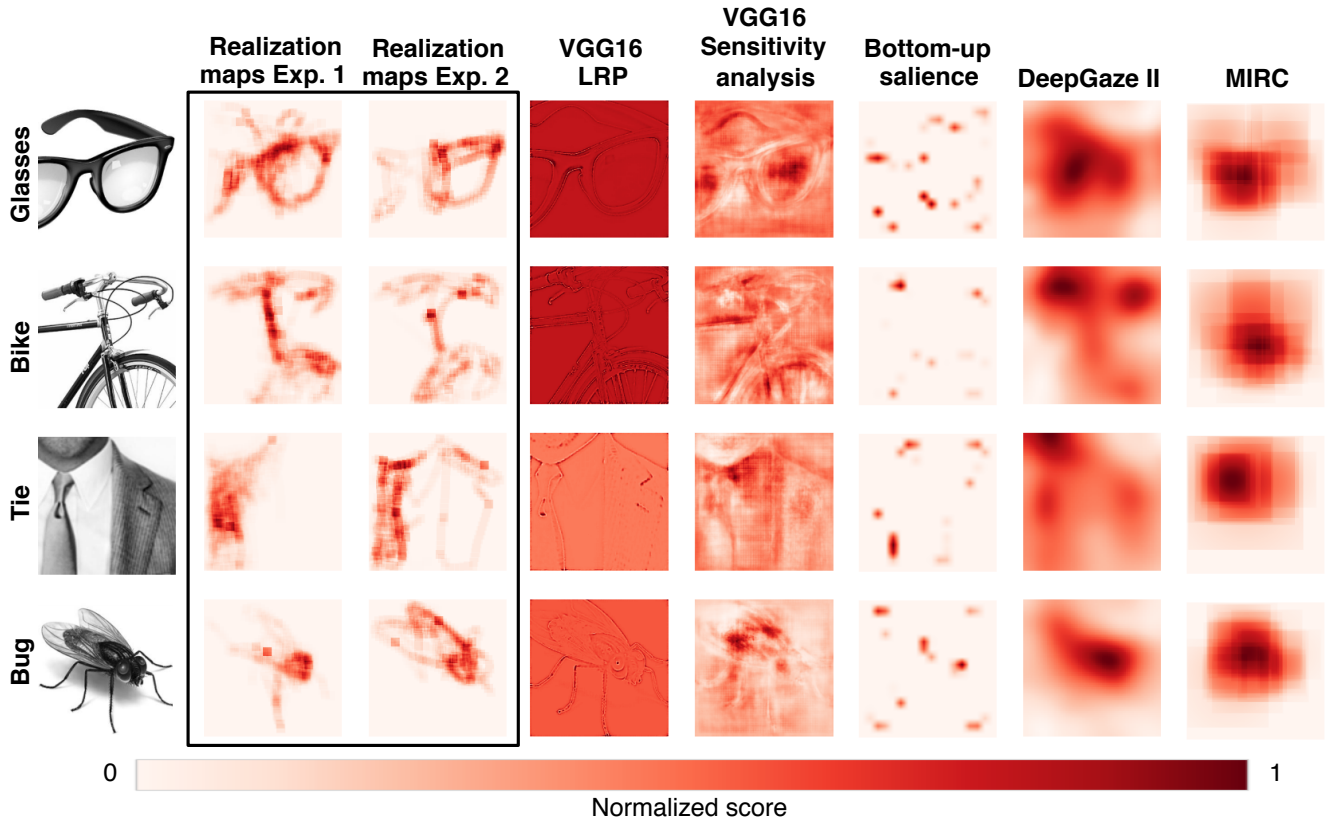


Figure 3. Heatmaps depicting object feature importance for humans and machines. From left to right: realization maps on images from Experiments 1 and 2 (the black box was added for emphasis), layer-wise relevance propagation maps (LRP) from VGG16, sensitivity analysis on VGG16, bottom-up salience, predicted salience from DeepGaze II, and MIRCs.

that penalized his team’s performance with the equivalent of 100 bubbles. Student and teacher switched roles after each round. The game was played for 110 rounds, with a different image each round. Each pair played on a random ordering of images.

Although participants could not communicate, we included features to make the game feel more collaborative. These were real-time notifications of what each player in the pair was doing at any point in time: clicking, typing, correct and incorrect responses, or thinking about what part of the image to reveal first.

We ran two versions of the *Clicktionary* game, referred to hereafter as Experiment 1 and Experiment 2. Both experiments included the ten images used in [31], for which we also had the MIRCs graciously provided by the authors. Beyond these, each experiment had participants judge a different set of 100 object images taken from the validation set of the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC; [23]). Images for each experiment were selected from 10 categories, 5 animate and 5 inanimate. For Experiment 1, we chose representative categories for animate objects: border collie, bald eagle, great white shark, and sorrel; and inanimate objects: airliner, school

bus, speedboat, sports car, and trailer truck. For Experiment 2, we chose the 5 animate and 5 inanimate categories that were the most difficult for VGG16 [29] to categorize. These were english foxhound, husky, miniature poodle, night snake, and polecat; cassette player, missile, screen, sunglasses, and water jug.

Games lasted about 20 minutes and participants were only allowed to play once. A total of 46 participants took part in Experiment 1, and 14 participants took part in Experiment 2. Participants were recruited through Amazon Mechanical Turk or from introductory cognitive science classes, and reimbursed approximately \$8.00/hr.

We created realization maps through a two-step procedure. First, the click path recorded between each pair of participants for an image was scored to emphasize pixels closer to recognition: $realization(x,y) = \frac{i}{j}$, where the realization score at a bubble centered at (x,y) is the current bubble index, i , divided by the total number j of bubbles revealed for the participant pair on the image. We used this formulation because it favors visual features that cause recognition. Bubbles that come right before recognition receive higher scores than those that do not, because we consider these features to reflect a tipping point in recognition. Our score

is also similar in spirit to the image features captured in MIRC’s [31].

Minimally reducible image configurations: Another source of information regarding important object features can be inferred from results reported in [31]. MIRC’s recovered for 10 images contain features that were critical for recognizing the object these images contained. We converted MIRC’s into heatmaps to support a systematic comparison with other measures of feature importance derived from both human participants and DCNs. This involved placing a 2D gaussian over the location of where each MIRC was discovered. The height of each MIRC gaussian was the inverse of the ratio of its pixel width and the MIRC with the smallest pixel width for that image. In other words, MIRC’s spanning large areas had smaller contributions to the MIRC heatmaps than MIRC’s spanning small areas. This scoring system was adopted to maximize our chances in identifying visual features contributing to MIRC’s. Each pixel in a MIRC heatmap was computed as the mean intensity value across these Gaussians. One of the images used in [31] depicted an image category not in ILSVRC 2012 and was excluded from analysis.

Models of object recognition: The primary objective of this study is to compare importance maps derived from human participants and DCNs. In support of this we produced DCN heatmaps of object importance for each of the images used in *Clicktionary*. This was done using VGG16, a variant of the popular VGG architecture [29]. We calculated heatmaps using a sensitivity analysis as done in [38] and a decision analysis (LRP) [1] for this model.

Models of attention: As controls, we considered the extent to which realization maps were explained through different models of attention. Theory holds that attention is driven by both bottom-up and top-down mechanisms [10]. We therefore compared realization maps to importance maps extracted from models for both kinds of attention, referred to hereafter as bottom-up saliency [10] and DeepGaze II [16].

4. Experiments

Realization maps: We systematically compared features used by humans to categorize representative objects to those used by DCNs. We use rank-order correlation throughout to measure these associations because realization maps are sparse and do not satisfy normality assumptions (no click map passed a Kolmogorov-Smirnov test for normality). All tests for significance used independent sample *t*-tests with two-tailed *p*-values.

Efficient

Inefficient

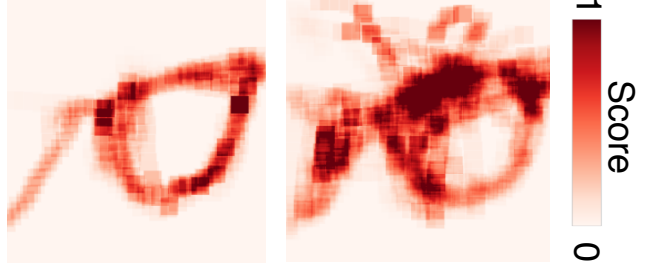


Figure 4. Realization map features vary according to how efficiently a participant pair recognized images. On the left, the mean realization map from pairs with above-median efficiency in recognizing glasses (i.e. faster recognition). On the right, the mean realization map from below-median pairs. The above image is representative of the typical differences between these groups.

We measured the inter-subject and inter-experiment consistency of the realization maps extracted from the *Clicktionary* game. Through two experiments, we observed that these realization maps are strongly stereotyped. Inter-subject consistency was computed using a randomization test ($n = 1000$) by correlating realization maps derived from participants randomly assigned into halves. There was a correlation of $\rho = 0.79$ ($p < 0.001$) in Experiment 1 and $\rho = 0.60$ ($p < 0.001$) in Experiment 2. We also measured consistency between participants in Experiment 1 and Experiment 2 on the MIRC images that both groups saw. Again, there was a strong correspondence between realization maps derived from these images for participants across the two experiments ($\rho = 0.55$, $p < 0.001$). Motivated by this agreement, all analyses hereafter use a pool of subjects from both experiments unless otherwise noted.

Despite the strong agreement we observed between *Clicktionary* participants, we found that realization maps were affected by player performance. Applying a median split to the number of bubbles it took before a student in a pair recognized an image revealed two qualitatively different realization maps (Fig. 4). Maps from the efficient group (i.e. number of bubbles is below median split) were sparser than those from the less efficient group. More work is needed to understand if these differences indicate increased noise from the latter group or belie different sets of visual features selected by each.

Humans versus machines: We separately compared importance maps from humans, DCNs, and attention models on images taken from ImageNet and the MIRC images. Strikingly, there was only a weak albeit significant relationship between realization maps and LRP derived from VGG16 ($\rho = 0.33$, $p < 0.001$; Fig. 5). By contrast, real-

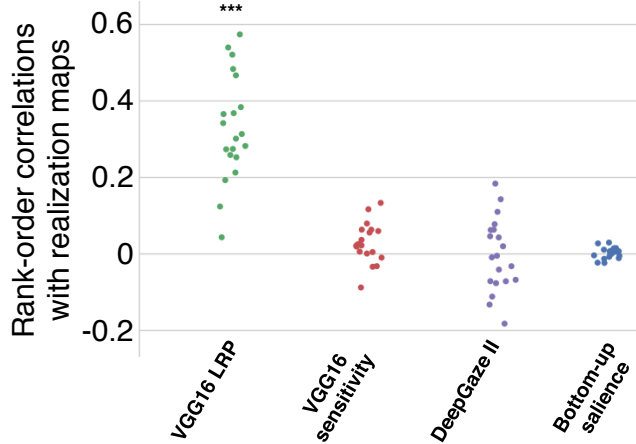


Figure 5. Correlations between human realization maps and feature importance maps derived from vision models. Each dot represents the mean correlation for individual image category. Independent samples t-tests measured deviation from 0. ***: $p < 0.001$

ization maps were not correlated with a sensitivity analysis performed on VGG16 ($\rho = 0.03$, n.s.). Realization maps were also not correlated with bottom-up salience ($\rho = 0.00$, n.s.) or eye fixation predicted by DeepGaze II ($\rho = 0.03$, n.s.; Fig. 5; also see supplementary information for representative heatmaps derived from human participants in Experiments 1 and 2.)

As a followup, we explored the similarity of human and machine feature importance maps on the images used to identify MIRC's [31]. This allowed us to include MIRC's in the comparison and understand their overlap with the feature importance maps discussed here. We found that DCNs, attention models, and realization maps qualitatively share at least some overlap with the regions occupied by MIRC's (Fig. 3) although the correlation was weak (Fig. 6; correlations depicted in the first column). As with ImageNet images, we found significant, albeit weak correspondence between realization maps and LRP ($\rho = 0.32$, $p < 0.001$; see Fig. 6 for details on each experiment). However, neither the sensitivity analysis nor the attention models were correlated with realization maps, and LRP was not correlated with the sensitivity analysis (Fig. 6).

These findings support our key assertion: visual strategies used by humans and DCNs during object recognition are not aligned. Realization maps capture mostly distinct information from DCNs, as measured by either LRP or a sensitivity analysis. There were even more pronounced differences between realization maps and feature importance predicted by attention algorithms². Given the motivation for our construction of realization maps, we expected a strong relationship with MIRC's. While the absence of this rela-

²The algorithm for predicting bottom-up salience maps was tuned to have qualitatively similar sparsity as the realization maps.

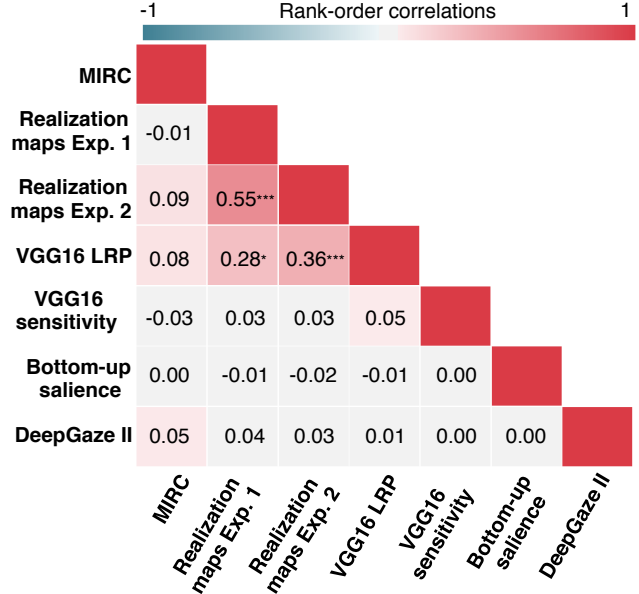


Figure 6. Mean pairwise correlations between realization maps derived from humans and DCNs using images from [31]. Cells are colored according to the strength and direction of association between maps.

tionship may be due to the coarseness of MIRC heatmaps or the small sample size used, the current evidence does not support realization maps as an alternate method for acquiring MIRC's.

Driving DCNs towards human realization: What is the explanation for the weak overlap between importance maps derived from human participants vs. DCNs? It could reflect basic mechanistic differences between humans and DCNs, such as a different number of processing layers or the presence/absence of top-down signals [6]. Another possibility is that it is indicative of possible biases and other limitations associated with the datasets and routines used for training DCNs. If this is the case, one would expect that cueing DCNs to attend to features that are diagnostic for human object recognition would improve object classification performance.

We tested this hypothesis on VGG16 and VGG19 top-1 classification accuracy on the ImageNet images used in *Clicktionary*. We chose top-1 classification (i.e. where classifier chance $\approx 1/1000$) because VGG has not yet reached ceiling accuracy (unlike top-5 classification). We compared the use of realization maps derived from human participants to salience maps derived from representative attention algorithms to cue DCN to attend to relevant object features during inference.

In each case, DCN activity for an image was modulated with a heatmap for that image: $Z \odot a$, where Z is each

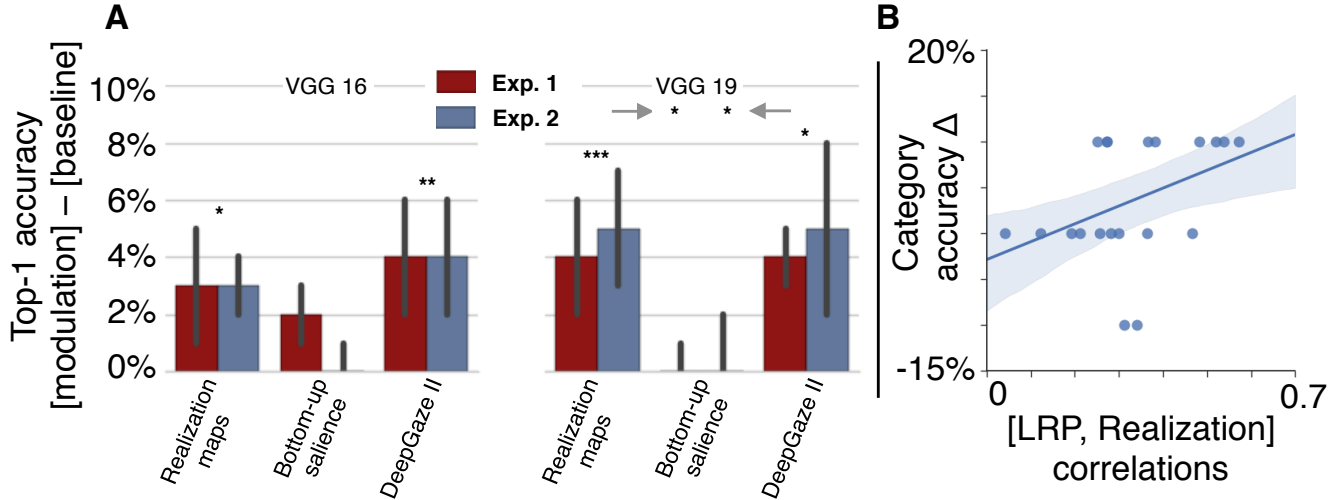


Figure 7. Modulating VGG16 and VGG19 activity with feature importance maps. **(A)** Realization maps and predicted fixations from Deep Gaze II but not salience maps improve VGG16 classification accuracy. A similar result was observed for VGG19. Note that here, 0% means no difference between an importance map modulated VGG and its baseline performance. Asterisks above bars are tests of classification accuracy improvement over baseline across Exp. 1 and Exp. 2. Asterisks attached to an arrow are pairwise comparisons. Error bars are S.E.M. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ **(B)** Realization map augmentation of VGG16 correlates with the association between its LRPs and realization maps.

model’s activity after the first layer of convolutions, and a is a heatmap scaled to the interval $[1, 2]$ and resized with bilinear interpolation to match the height and width of Z . Each model’s top-1 accuracy was recorded after modulation. Rescaling heatmaps was crucial as it ensured that there was amplification of important features instead of attenuation, which can lead to artifacts in the activations that are pathologically interpreted by downstream layers.

We found that modulating VGG16 and VGG19 activity with realization maps yielded a small but statistically significant improvement in top-1 accuracy. VGG16 modulated with realization maps was on average 3.49 percentage points better in top-1 accuracy than the baseline performance of VGG16 (Fig. 7A; left panel, realization map deviation from 0: $p = 0.025$). Similarly, VGG19 was on average 4.00 percentage points better than the baseline performance of VGG19 (Fig. 7A; right panel, realization map deviation from 0: $p = 0.004$). Note that there was a larger magnitude of improvement for images in Experiment 2 than in Experiment 1, likely reflecting that the latter experiment’s images were more difficult to recognize, as expected.

These gains mean that DCNs benefit from focusing on features that trigger object recognition in humans. Supporting this, the gains in accuracy we observed for each object category were associated with the fisher-transformed correlation coefficients between realization maps and LRPs ($\rho = 0.41$, $p = 0.069$; Fig. 7B). This means that in our current implementation DCNs received the greatest benefit from realization maps when they were already attuned to

the features emphasized by those maps.

DCN accuracy was not significantly different from 0 when it was modulated with bottom-up salience. VGG19 modulated with bottom-up salience was also significantly less accurate than when it was modulated with realization maps or predicted fixations from DeepGaze II (Fig. 7A; right panel, realization map deviation from bottom-up salience: $p = 0.023$; DeepGaze II deviation from bottom-up salience: $p = 0.026$). Although fixations predicted by DeepGaze II lead to a similar improvement in classification accuracy as realization maps, these gains were mostly accrued on different object categories. There was weak overlap between the categories improved by realization maps and those improved by DeepGaze II ($\rho = 0.36$, $p = 0.022$). We believe that future work incorporating feature maps into training (instead of only during inference, as we do here) can provide more clarity to these differences. Specifically, realization maps are qualitatively sparser than DeepGaze II, which could provide important gains for the efficiency of training.

5. Discussion

Clicktionary is a novel approach to estimate feature importance maps derived from humans. The proposed method overcomes some of the main limitations from existing psychophysical methods including reverse correlation and other image classification methods (see [19] for a review). We have described what is, to our knowledge, the first systematic study of feature importance maps derived

from human participants using natural images over multiple object categories.

We found little or no overlap between realization maps and image salience maps predicted by attention models. This suggests that realization maps reflect computational mechanisms that are somewhat distinct from those that guide attention and eye movements. It is important to note that realization maps are of course generated through a process that necessarily depends on attention: Teachers' bubbles reflect a continuous ranking of the importance of image features for recognition. However, this is distinct from typical methods for measuring (or predicting) attention in two ways. First, salience is usually measured passively or for a search task, whereas relevance maps take advantage of the teacher's ability to select important features, trading salience for feature diagnosticity. Second, the interplay of teacher and student identifies the point at which visible features become sufficient to trigger recognition. Our current method for visualizing realization maps across participant pairs remains relatively simple – potentially disregarding important information through averaging. We expect that future work in developing a more nuanced approach for measuring realization maps from *Clicktionary* will prove useful for further characterizing visual strategies for both humans and DCNs.

We found a weak association between realization maps and importance maps in VGG16 derived from LRP. This indicates that there exists at least some overlap between visual features used by humans and DCNs for object categorization. However, the magnitude of this association was around half of what it was between Clicktionary participants, suggesting that the visual representations used by DCNs and humans are still meaningfully different.

Surprisingly, emphasizing realization maps in DCNs during inference nonetheless improved classification accuracy. In addition, the strength of this augmentation was predicted by the amount of overlap between realization maps and LRPs. These results therefore indicate that driving DCNs towards human recognition is beneficial, and realization maps can support this optimization.

One way of achieving this is to train computer vision models that can predict the location of realization maps in images. Recent work has demonstrated support for this approach: Models that leverage DCN features achieve state-of-the-art accuracy in predicting eye fixations [16]. We hypothesize that models of this ilk will have similar success in predicting realization maps. Unfortunately, while the experiments reported here are significantly larger-scale than any others for measuring visual feature sensitivity in humans, they do not yield enough data for training these kinds of models.

To rectify this, we present clickme.ai, a web-based application for gathering recognition maps at scale (Fig. 8).

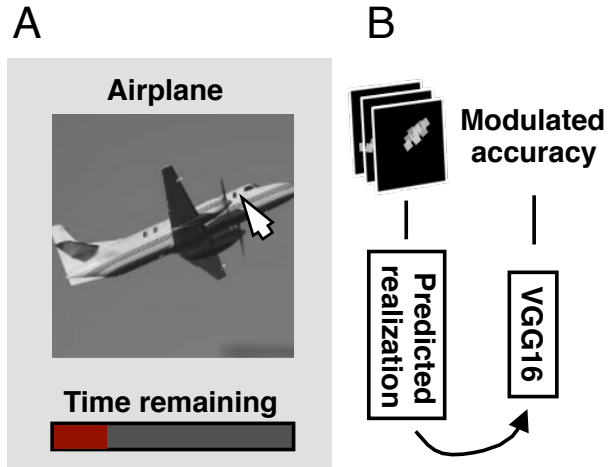


Figure 8. A depiction of the interface at clickme.ai for large-scale acquisition of realization maps, and continuous measurement of its impact on DCN performance. Participants draw bubbles over important regions in an object image, as a DCN tries to recognize it before a time-limit elapses. After accruing many of these realization maps, the application trains a model for predicting realization maps in images. These maps are applied to VGG16 during training to drive it towards human perception.

Upon arriving to the website, players are presented with an object image and asked to play the role of teacher, using the mouse to place bubbles on its most critical features. The student in this case is a DCN, attempting to classify the object from the bubbled image regions within a time-limit. Players compete for prizes to get the DCN to recognize the most images.

After accruing a significant amount of these human-DCN realization maps, clickme.ai triggers a two-step training sequence. First, a DCN designed for predicting eye fixations is trained to predict the location of participants' bubbles [3]. Second, a VGG16 is cued to hot image regions in the predicted realization maps during training for object recognition. The resulting accuracy is posted to clickme.ai to provide a running tally of the benefit of predicted realization maps on accuracy over baseline, which has no modulation. Thus, in contrast to popular adversarial learning methods that have achieved spectacular success in recent years (e.g., [7]), clickme.ai represents an attempt at “cooperative training”, in which a human-in-the-loop trains an auxiliary system that will yield more effective models for object recognition.

Overall, *Clicktionary* makes significant contributions to our understanding of biological vision and reveals a significant gap between feature importance for humans and machines. These findings will inspire new directions in DCN research and narrow the gap between biological and computational vision.

Acknowledgments

The original idea for Clicktionary was suggested by Prof. Todd Zickler (Harvard University). We are also thankful to Matthias Kümmerer for running DeepGaze on our images and for Danny Harari and Shimon Ullman for providing MIRC stimuli used in [31]. We are also indebted to Junkyung Kim and Matthew Ricci for comments on the manuscript. This work was supported by NSF early career award (IIS-1252951) and DARPA young faculty award (N66001-14-1-4037).

References

- [1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. *PLoS One*, 10(7):e0130140, 10 July 2015. 5
- [2] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 5 Aug. 2010. 3
- [3] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *23rd International Conference on Pattern Recognition (ICPR)*, 2016. 8
- [4] A. Das, H. Agrawal, C. Lawrence Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? 11 June 2016. 3
- [5] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1089–1097, 2015. 2
- [6] S. Eberhardt, J. Cader, and T. Serre. How deep is the feature analysis underlying rapid visual categorization? In *Neural Information Processing Systems*, 2016. 1, 6
- [7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 20 Dec. 2014. 8
- [8] M. R. Greene, T. Liu, and J. M. Wolfe. Reconsidering yabus: a failure to predict observers’ task from eye movement patterns. *Vision Res.*, 62:1–8, 1 June 2012. 2
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016. 1
- [10] L. Itti and C. Koch. Computational modelling of visual attention. *Nat. Rev. Neurosci.*, 2(3):194–203, 2001. 5
- [11] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, June 2015. 2
- [12] Y. V. Jiang, B.-Y. Won, and K. M. Swallow. First saccadic eye movement reveals persistent attentional guidance by implicit learning. *J. Exp. Psychol. Hum. Percept. Perform.*, 40(3):1161–1173, June 2014. 2
- [13] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *CVPR*, 2009. 2
- [14] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier. Humans and deep networks largely agree on which kinds of variation make object recognition harder. 21 Apr. 2016. 1
- [15] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung, and EyeWisers. Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500):331–336, 15 May 2014. 3
- [16] M. Kümmerer, T. S. A. Wallis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. 5 Oct. 2016. 5, 8
- [17] B. M. Lake, W. Zaremba, R. Fergus, and T. M. Gureckis. Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2015. 2
- [18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. pages 740–755. 2
- [19] R. F. Murray. Classification images: A review. *J. Vis.*, 11(5):2–, 1 Jan. 2011. 7
- [20] K. J. Nielsen, N. K. Logothetis, and G. Rainer. Object features used by humans and monkeys to identify rotated shapes. *J. Vis.*, 8(2):9.1–15, 22 Feb. 2008. 2
- [21] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. pages 361–376. 3
- [22] J. C. Peterson, J. T. Abbott, and T. L. Griffiths. Adapting deep network features to capture psychological representations. 6 Aug. 2016. 2
- [23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *CoRR*, abs/1409.0:43, 1 Sept. 2014. 4
- [24] B. Saleh, A. Elgammal, and J. Feldman. The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks. 9 Feb. 2016. 1
- [25] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1679–1686, Aug. 2014. 3
- [26] P. G. Schyns, L. Bonnar, and F. Gosselin. Show me the features! understanding recognition from the use of visual information. *Psychol. Sci.*, 13(5):402–409, Sept. 2002. 2
- [27] K. Shanmuga Vadivel, T. Ngo, M. Eckstein, and B. S. Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3250, 2015. 3
- [28] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014. 2
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for Large-Scale image recognition. *Intl. Conf. on Learning Representations (ICLR)*, pages 1–14, 2015. 4, 5

- [30] J. Steggink and C. G. M. Snoek. Adding semantics to image-region annotations with the Name-It-Game. *Multimedia Systems*, 17(5):367–378, 9 Dec. 2010. 3
- [31] S. Ullman, L. Assif, E. Fetaya, and D. Harari. Atoms of recognition in human and computer vision. pages 1–6, 2016. 1, 4, 5, 6, 9
- [32] E. Vig, M. Dorr, and D. Cox. Space-Variant descriptor sampling for action recognition based on saliency and eye movements. pages 84–97. 3
- [33] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’04, pages 319–326, New York, NY, USA, 2004. ACM. 2
- [34] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’06, pages 55–64, New York, NY, USA, 2006. ACM. 2
- [35] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, 23 Feb. 2016. 1
- [36] A. L. Yarbus. *Eye Movements and Vision*. Plenum press, New York, 1967. 2
- [37] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *arXiv preprint arXiv:1311.2901*, pages 818–833, 2013. 2
- [38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene CNNs. 22 Dec. 2014. 2, 5

Supplementary information

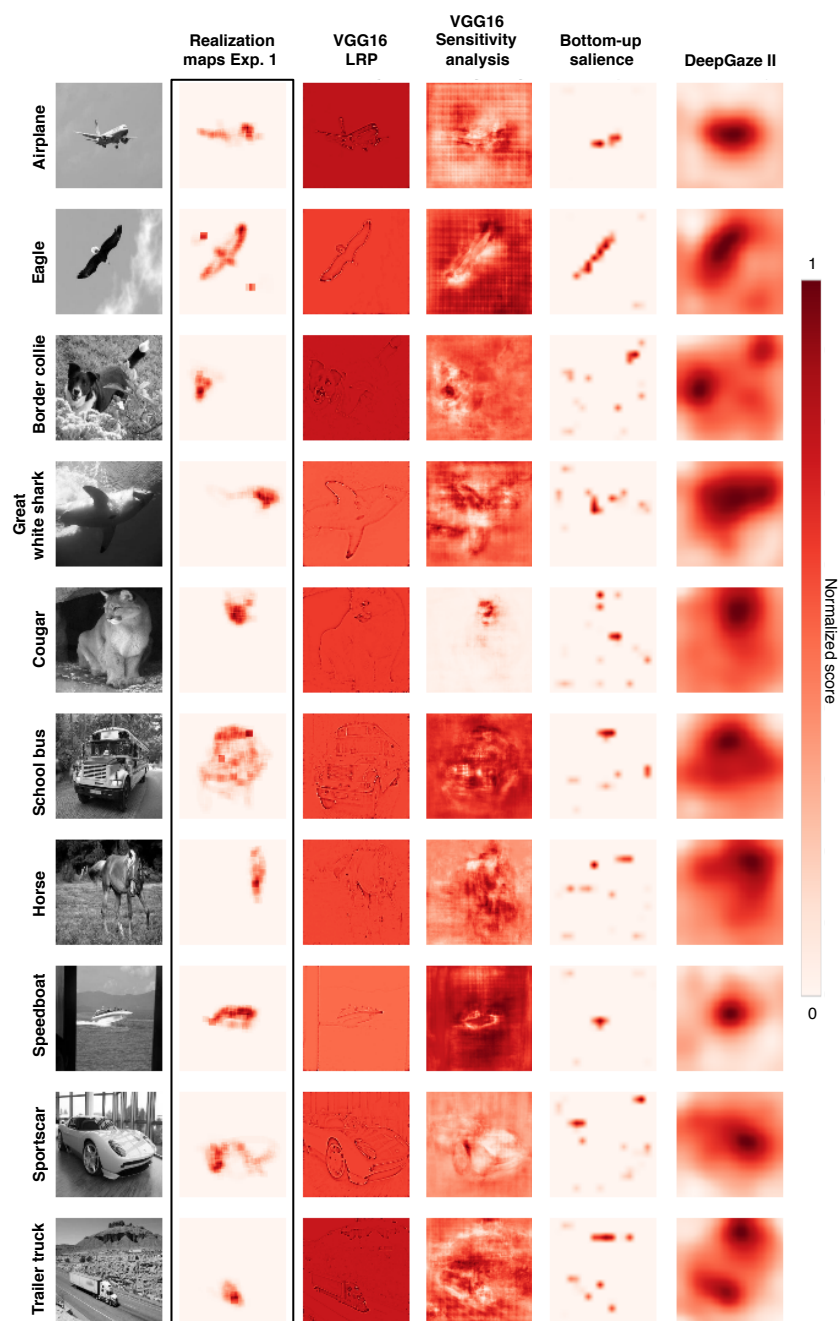


Figure S1. Feature importance maps for images from Experiment 1.

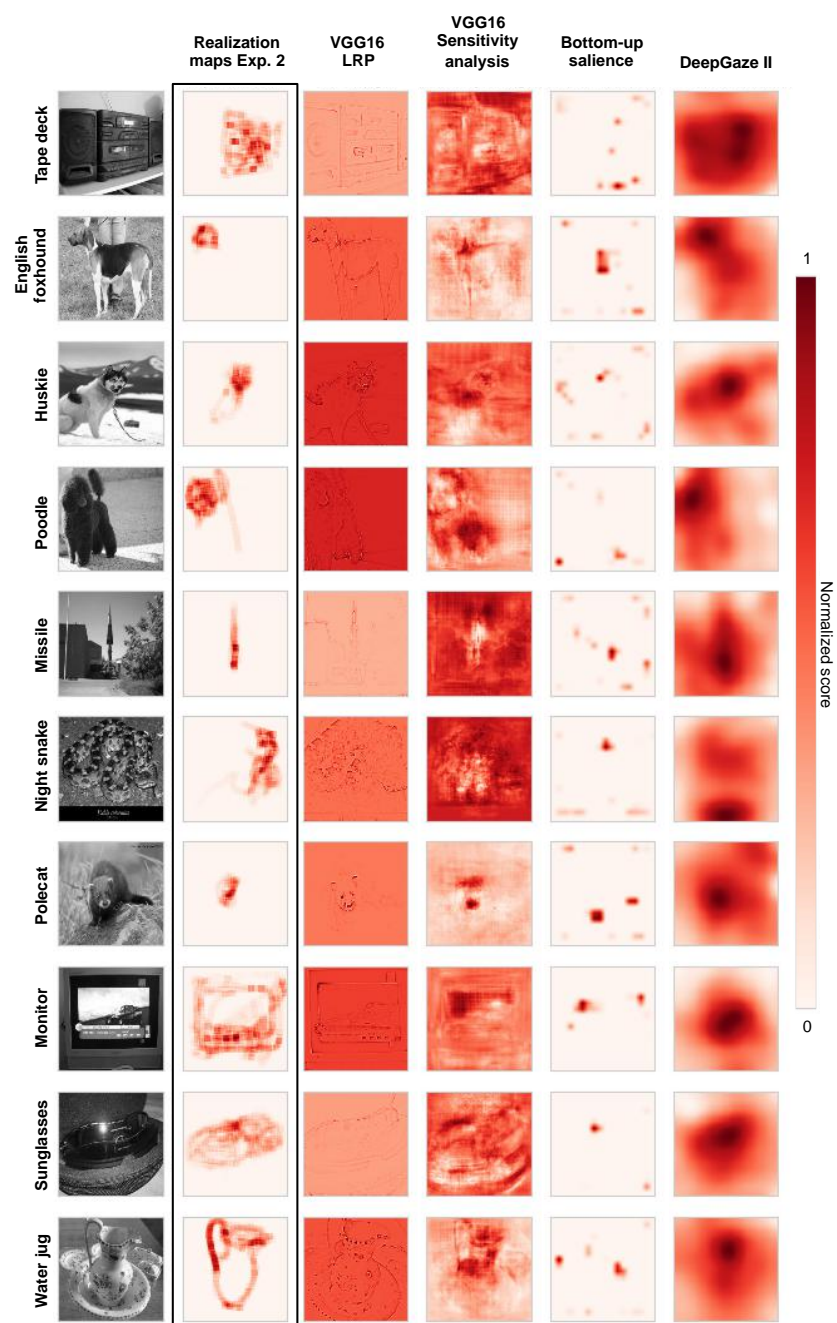


Figure S2. Feature importance maps for images from Experiment 2.